MANUAL AND AUTOMATIC ALIGNEMENT OF PAGES

RELATED APPLICATIONS

This application claims the benefit of the filing date Provisional Application Serial Number 60/453,937 filed 03/12/03 entitled "MANUAL AND AUTOMATIC ALIGNEMENT OF PAGES".

BACKGROUND

This invention relates to printing books from scanned pages and particularly concerns aligning the pages in the printed book.

A Portable Document Format (PDF) file is a self-contained cross-platform document. It is a file that will look the same on the screen and in print, regardless of what kind of computer or printer someone is using and regardless of what software package was originally used to create it. A PDF file contains the complete formatting of the original document, including font information and images. A user may compress the PDF file and this allows complex information to be stored, transmitted and downloaded efficiently. The PDF file format was developed by Adobe Systems. PDF captures formatting information from a variety of desktop publishing applications, making it possible to send formatted documents and have them appear on the recipient's monitor or printer as they were intended. The PDF file format captures all the elements of a printed document as an electronic image that can be viewed, navigated, or printed. In the PDF format, all the text information is still available as text, all graphics are still stored as vector graphics, and images are still images. PDF files are more than images of documents. PDF files can have embedded resources, like type fonts so that these resources are available at any viewing location.

PDF files are especially useful for documents such as magazine articles, product brochures, or flyers where a publisher wants to preserve the original graphic appearance online. A PDF file contains one or more page images. A viewer may zoom in on or out from any page and may move forward and backward. Because PDF files are cross platform files, they are widely used in the printing industry. A customer often brings to a printer documents

that are in different formats and asks the printer to arrange the documents into a book. Other times a customer has a hard copy of book that he wants to reproduce and the easiest method of reproduction is to copy the pages on a photocopy machine and sort the pages into books.

To scan a book and therefore convert it to an electronic version that can be printed on a production printer, the book is cut at the spine and then scanned page by page. This creates originals that no longer contain perfectly aligned page images. When the raw scan data is printed, the pages appear to "jump" around when one browses through the book. Because of that, there is a need to either automatically or manually align the pages prior to printing. This invention assumes a PDF based workflow. It is also assumed that the application providing these features is implemented as an Acrobat plug-in. All image pre-processing like deskew or despeckle are performed before the pages are aligned in accordance with one of the following procedures. This pre-processing may automatically be initiated by the alignment function before the actual page alignment.

SUMMARY

The invention provides a computer program for operating a printing machine. It includes a method of operating the machine as well as the machine that operates the method of the program. In particular, the invention provides a method, apparatus and program for aligning pages in a book and printing the book. With the invention a user scans pages of a book into a controller or computer where the pages are converted into a file that is independent of the platform that created the documents. In the preferred embodiment the computer converts the scanned pages into portable document formatted (PDF) pages. Each PDF page comprises content areas of text or graphics or both and non-content areas surrounding the content areas. The program determines a bounding box around the content areas for placement purposes and to determine the size (and position) of the content area. As such the page is temporarily cropped in order to asses its content area and align its content area with other pages. As soon as this information is available, the cropped image is discarded again. The original image is moved by the amount determined through this operation. During automatic alignment, the program operates on the files to temporarily crop or remove peripheral, non-content areas and generate cropped PDF pages of the content areas on the pages of the book. The cropped PDF files normally includes one image per page although it may encompass text and graphics. The program may operate in

automatic or manual mode. In other words, the program processes the PDF formatted document to define the minimal content areas that include all of the information (text, graphics and images) on the page. This is called a "cropped" image for purposes of explaining how the program works. No portion of the document is actually removed. In either automatic or manual mode, either the program or the user selects a feature of the cropped PDF pages common to all the PDF pages. The rest of the pages are aligned to the selected feature and the book is printed.

The manual method lets the user select one or more group(s) or subset(s) of PDF pages. Then the user previews the selection and chooses a standard page and a feature on the standard page that is suitable for aligning to other pages. The user places a first cursor on the standard feature of the content of the previewed PDF page. The other pages are each previewed and each subsequent page displays the first cursor in the position selected on the standard page. The program places a second cursor on the second and subsequent pages and lets the user move the image as a block to align the cursors to one another and thereby manually align the pages.

DRAWINGS

Fig. 1 shows three exemplary pages of a scanned book.

Fig. 2 shows the three exemplary pages cropped after scanning.

Fig. 3 shows the three pages of Fig. 1 automatically aligned.

Fig. 4 shows three exemplary pages with compound images.

Fig. 5 shows the three pages of Fig. 4 aligned.

Fig. 6 shows three exemplary pages undergoing manual alignment.

Fig. 7 is a schematic diagram of a scanner and printer equipped with the computer program of the invention.

Fig. 8 is a flow diagram of the steps performed by the invention.

DETAILED DESCRIPTION

Automatic Alignment

All scanned pages usually contain white space around the actual page content. The alignment program removes the white space by cropping the page. In operation, the scanned pages are converted into PDF pages and at least one reference point (e.g. upper left corner) is always placed at the same position. Removing the white space is relatively easy for true scanned PDF pages and is readily implemented by one skilled in the art. The cropped pages contain only one PDF object: The scanned image. Only ordinary skill in the art is required for converting compound pages that contain more than one image object, or other PDF objects, into cropped pages with one object in a bounding box.

Turning to Fig. 1, there are three original scanned pages, P1ORG, P2ORG and P3ORG. Each page includes text, T1, T2, T3, respectively. The text on each page is surrounded by a border area of white or blank space, W1, W2, W3, respectively. On page P1ORG the white space W1 is indicated in cross hatch.

Scanned Pages

For scanned pages, the PDF pages will contain only one PDF object: One image containing the full page image. See Fig. 2. For these pages, the automatic alignment is relatively easy to perform. The pages are pre-processed with an "auto-crop" function, which automatically removes the white border around the page content. When the pages are cropped, the blank, border areas W1, W2, and W3 are removed and the pages are limited to their respective content areas inside transient bounding boxes B1-B3. The cropped images are essentially only the text portions, T1, T2, and T3. See Fig. 2. Each block of text T1-T3 is treated as a single object.

After auto-crop, the user defines one reference point (e.g. upper left corner). The program moves the page images so that the reference point on the individual pages line up with each other. See Fig. 3 where the pages are automatically aligned as shown for pages P1AA, P2AA, and P3AA. An alternate embodiment allows the program to vote the reference point. In other words, the user need not define a reference point. Instead, the software measures the sizes of all the images T1-T3 to identify the largest page and then

positions the largest page according to rules that were either defined by the user (e.g. left edge), or internal to the software. All other pages are placed according to this largest page.

Compound Pages

Books often include pages that contain text and graphics. See for example the original pages in Fig. 4 where images I1, I2, and I3 appear on the pages. If the page layout contains multiple PDF objects, in order to layout all pages in a uniform way the software makes a bounding box for all pages: The bounding box is the smallest theoretical box B1, B2, B3, that fully encloses all objects on a page. The bounding box is established by counting pixels in the scanned image. There are many ways of scanning a document to identify blocks of text and graphics and then designate those areas as objects. In this alignment program, all blocks of text and graphics are combined into one block or area that includes all text and graphics. The program creates, in effect, a large rectangular bounding box that encloses all text and graphics on a page and excludes border white space. This is done by determining the bounding boxes of all objects on the page, and then combining these bounding boxes to a resulting bounding box by iterating over all bounding boxes and comparing the bounding box of an individual object with the resulting bounding box. If the individual bounding box's lower left corner is left of the resulting bounding box's lower left corner, then the x component of the lower left corner of the resulting bounding box will be assigned the x component of the lower left corner of the individual bounding box and so on for all of lower left x, lower left y, upper right x, upper right y. Once this bounding box is established for all pages, the process continues as described in the previous section ("Scanned Pages"). In operation, the program creates a transient bounding box that includes the image and text on each page and automatically aligns the bounded objects. See the result in Fig. 5.

Page Size Comparison

The software provides a mechanism to validate the pages sizes (bounding boxes) by grouping them into similar page size groups (or subsets). If all pages have approximately the same size, they can be aligned without any manual intervention. If the software

however finds a few pages that have different bounding boxes (e.g. the start of a chapter where the first page is not completely filled with text), the user can be warned. In addition to this, the software may be able to align the pages that fall into one page size group in a consistent way. The user can define new alignment rules for a group of pages with comparable page sizes.

Preview

The software provides a preview function that displays the results of the cropping alignment for one selected page without applying the changes to the actual document. The preview function either opens a new window to display the cropped aligned page view, or creates a new (temporary) document containing just the one cropped aligned page. This allows the operator to zoom into the page image to verify the correctness of the cropping operation.

Manual Alignment

For the manual alignment of the scanned pages, the operator reviews all pages and defines one page as a "standard" for all another alignment operations. Turning to Fig. 6, assume the operator selects PM1 as the standard. The operator will mark one position on this standard page that will have an equivalent position on all other pages (e.g. upper right corner of the text, left most point of a horizontal rule below the heading, left lower corner of a page number). In Fig. 6 the operator has chosen the three asterisks that appear as a footer in the center of each page and place a first cursor C1 at the end of the first asterisk. On the first page the user may move the entire text box T1 to its desired position. Once positioned, the user locates a first marker cursor C1 (e.g. a circle marker) to a position proximate the asterisks.

The first cursor circle marker C1 marks the correct "zero position" on the first page. Each subsequent page has a C1 cursor marker that shows where the zero position would be without any correction. A hexagon marker cursor C2 on the second and subsequent pages allows the user to move the "zero point" to a new location to define a point on the page which should be moved to the zero position specified by the first cursor by shifting the entire contents of the page accordingly. The names and shapes of the markers used in this

description are only samples; the actual implementation of the software may use different names or shapes.

Printing Mode /Scanning Mode

Books are usually printed in duplex mode and they are often scanned in duplex mode. Usually the left and right pages of a book have a slightly different layout (e.g. a left page has the page number at the left edge, a right page has it at the right edge). The program has an option to process all front or odd pages differently from all back or even pages. Because the document was scanned in duplex mode, it can be assumed that all front pages have odd page numbers, and all back pages have even page numbers. Therefore the program provides instructions for a computer controlled scanner and printer to process all odd page numbers with one reference setting, and all even page numbers with a different reference setting.

The user may select two different reference points, one for even and one for odd page numbers. When processing the pages, the program presents all odd numbered pages first, and after that all even numbered pages next. The advantage of this over presenting all pages in their original order is that the user could concentrate on one reference point, and would not have to jump from the left edge of a scan to the right edge for the following page. It may also be possible to further limit the set of pages to be processed by creating a page selection in Acrobat's thumbnail menu. The software would then only present the selected pages to the user for a manual page alignment.

Example

Figure 7 shows a schematic layout of printer 14 equipped with the computer program of the invention. The figure and the following description are generic for printers and are not limited to the particular details provided herein. Those skilled in the art understand that printers may have many different configurations. As such, the following description is provided to enable one skilled in the art to understand the environment in which the page alignment feature of the invention is employed. Exemplary printers include the Digimaster.TM. Digital High Volume Printer manufactured by Heidelberg Digital, L.L.C., located in Rochester, N.Y. and the NexPress TM color printer manufactured by NexPress, Corporation, located in Rochester, N.Y.

Scanner 10 scans a document 12. The scanner is preferably a sheet fed scanner that rapidly scans the pages of a book that has been disassembled. The scanned images are received by a controller 20 that includes a central processing unit 22 and one or more memory units 24. The memory unit(s) include random access memory and read only memory for holding data, system programs, and application programs including (and not limited to) Adobe Acrobat and the alignment program of the invention. The programs run on the CPU 22 are under control of an operator who has a display terminal 26 and input devices such as a keyboard 28 and a mouse 30.

Printer 14 has media input bins (not shown) that hold the paper for printing a book. After the alignment program is run and the print job is otherwise ready for release, the printer prints the aligned images onto the media and discharges the printed book. Other downstream options may include a cover inserter and binder or the binding may be as simple as stapling the pages together.

In operation, the book is separated into its individual pages as shown in step 100 of Fig. 8. Then the scanner 12 scans each page (step 101) and the scanned images are stored in a memory unit 24 of the controller 20. In the preferred embodiment, the CPU runs Adobe Acrobat and converts the scanned images of the pages into PDF files. The invention program then crops the PDF files in step 102 to provide the cropped bounding shown in Fig.2. The user then operates the alignment program in step 103 to select whether the book will be aligned to one standard (simplex) to two standards (duplex) such as one for odd pages and another for even pages, or multiple standards, one for each page grouping. In step 105 the user selects automatic or manual alignment for each grouping. If automatic alignment is selected, then the user may select a standard or let the program determine a reference point for each grouping in a step 106. In a step 107 the program compares the pages in the grouping and the program identifies the largest page and aligns all the pages (or at least all the pages in the grouping) to the page with the largest bounding box in each grouping. If the user selects manual alignment (step 108) then he further selects a standard for that process (step 109). The user may preview (step 112) each page on the display 26 to see if manual alignment is needed or if the initial scanned PDF page is sufficiently close to the standard to be acceptable. After alignment is completed, the book is printed (step 110 and/or stored to disk for a later printing (or re-printing).